BIOINFORMATICS ORIGINAL PAPER

Vol. 29 no. 2 2013, pages 238–245
doi:10.1093/bioinformatics/bts670

Systems biology

Advance Access publication November 17, 2012

Drug–target interaction prediction by learning from local information and neighbors

Jian-Ping Mei[1,*], Chee-Keong Kwoh[1], Peng Yang[1], Xiao-Li Li[1,2] and Jie Zheng[1]

[1]Bioinformatics Research Centre, School of Computer Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore 639798, Singapore and [2]Institute for Infocomm Research, A*Star, 1 Fusionopolis Way #21-01 Connexis, Singapore 138632, Singapore

Associate Editor: Trey Ideker

ABSTRACT

Motivation: *In silico* methods provide efficient ways to predict possible interactions between drugs and targets. Supervised learning approach, bipartite local model (BLM), has recently been shown to be effective in prediction of drug–target interactions. However, for drug-candidate compounds or target-candidate proteins that currently have no known interactions available, its pure 'local' model is not able to be learned and hence BLM may fail to make correct prediction when involving such kind of *new candidates*.

Results: We present a simple procedure called neighbor-based interaction-profile inferring (NII) and integrate it into the existing BLM method to handle the *new candidate* problem. Specifically, the inferred interaction profile is treated as label information and is used for model learning of new candidates. This functionality is particularly important in practice to find targets for new drug-candidate compounds and identify targeting drugs for new target-candidate proteins. Consistent good performance of the new BLM–NII approach has been observed in the experiment for the prediction of interactions between drugs and four categories of target proteins. Especially for nuclear receptors, BLM–NII achieves the most significant improvement as this dataset contains many drugs/targets with no interactions in the cross-validation. This demonstrates the effectiveness of the NII strategy and also shows the great potential of BLM–NII for prediction of compound–protein interactions.

Contact: jpmei@ntu.edu.sg

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 2, 2012; revised on October 15, 2012; accepted on November 12, 2012

1 INTRODUCTION

Identification of interactions between drugs/compounds and protein targets is an important part of the drug discovery pipeline. The great advances in molecular medicine and the human genome project provide more opportunities to discover unknown associations in the compound–protein interaction network. The newly discovered interactions are helpful for discovering new drugs by screening candidate compounds and also may help understand the causes of side effects of existing drugs. Since experimental way to determine drug–target interactions is costly and time-consuming, *in silico* prediction becomes a potential complement that provides useful information in an efficient way.

Generally, the prediction performance is decided by both the data used and the particular analysis method that is applied to. An intuitive and straightforward way to identify new targets for a drug is to compare the candidate proteins with those existing targets of that drug. Different results may be obtained depending on which perspective the comparison is made with respect to. Keiser *et al*. (2009) compare targets based on the chemical structure of ligands that bind to them. As reviewed in Haupt and Schroeder (2011), the structure of binding sites is another important way to compare proteins or to measure the similarity between proteins. Although binding site is an effective measure for identification of new targets, the structures of binding site are only available for a small set of proteins, of which the 3D structures are known. To be able to consider more proteins, amino acid sequence may be used as it is available for most proteins. Similarly, to identify new targeting compounds for a specific target, comparison is made on the compound side or drug side with respect to chemical structures (Laggner *et al*., 2012; Martin *et al*., 2002), side effects (Campillos *et al*., 2008) or other possible measurements of drug.

More sophisticated statistical and machine learning methods have been developed recently for prediction of genome-wide drug–target interactions. In He *et al*. (2010) and Perlman *et al*. (2011), multiple groups of drug-related features and protein-related features have been extracted to describe each drug–target pair. After feature selection, a certain classifier is used to predict whether a given pair is interacting or not. Yamanishi *et al*. (2008) proposed a supervised bipartite graph learning approach. In this approach, the chemical space and the geometric space are mapped into a unified space so that those interacting drugs and targets are close to each other while those non-interacting drugs and targets are far away from each other. By mapping the query pair of drug and target to that space with the learned mapping function, the probability of interaction between them is then calculated as their closeness in the mapped space. Another method called the weighted profile method was also given in Yamanishi *et al*. (2008). For a query drug, the weighted profile method assigns a probability of interaction to the query target based on how the neighbors of this drug interact with this target. Basically, weighted profile is a nearest-neighbor approach and it is called drug-based/target-based similarity inference in Cheng *et al*. (2012). Other than

*To whom correspondence should be addressed.

inferring interactions from the drug similarity or target similarity, network-based inference was also studied in Cheng *et al*. (2012), which infers or predicts drug–target interactions based on the topology of the known interaction network. Different from the work in Cheng *et al*. (2012), which makes use of the drug similarity, target similarity and network-based similarity separately, Chen *et al*. (2012) apply random walk on a heterogeneous network constructed with these three types of similarities. Another promising approach is the bipartite local model (BLM) approach. Bleakley and Yamanishi (2009) showed that the ensemble of independent drug-based prediction and target-based prediction with supervised learning performs much better than only using each single type of prediction. The BLM method has been further studied and improved in Xia *et al*. (2010) and Laarhoven *et al*. (2011). The main differences of these three methods include the drug–drug and target–target similarities, the classifiers and the way used to combine the drug-based and target-based interaction probabilities. In Xia *et al*. (2010), semi-supervised approach is used instead of supervised approach for local model learning; while Laarhoven *et al*. (2011) found that using only the kernel based on the topology of the known interaction network is able to obtain a very good performance.

In the existing framework of BLM, the model for the query drug or target is learned based on local information, i.e. its own interaction profile. Despite a good performance, BLM has limitations. It is unable to learn without training data and hence is not able to provide a reasonable prediction for drug/target candidates that are currently new. Here, a drug-candidate compound is *new* if it does not have any known targets, and a target-candidate protein is *new* if it is not targeted by any drugs/compounds. We call this the *new candidate* problem of BLM. Since a large number of compounds and proteins, which are possible drug candidates and target candidates, respectively, are *new*, in this study, we focus on handling the *new candidate* problem by proposing an improved version of BLM called BLM with neighbor-based interaction-profile inferring (BLM–NII). The NII procedure is developed to incorporate the capacity of learning from neighbors into the original BLM method. More specifically, when the query involves a new drug/target candidate, we first derive the initial weighted interactions for the new candidate from its neighbors' interaction profiles, and then use the inferred interactions as label information to train the model. In general, *neighbors* refer to compounds/proteins that have large similarities to the query compound/protein.

The presented NII idea happen to be similar to the weighted profile method in some sense. However, our BLM–NII method is substantially different from the weighted profile method in the following aspects. In BLM–NII, the derived interaction profile is used as label information to train the local model or the classifier, while in the weighted profile method, the derived weighted interaction is directly used as the final predicted interaction probability. Moreover, in BLM–NII, the NII procedure is integrated into the BLM framework where a certain classifier plays the main role in model learning, and NII is activated only for new drug/target candidates; while in the weighted profile method, there is no other classifier and the procedure of deriving the weighted profile acts as a classification process, which is applied for any drug/target candidates. To sum up, the BLM–NII is an enhanced BLM method, and it is different from the weighted profile method, which is a nearest-neighbor approach. Our experimental results show that BLM–NII performs much better than the weighted profile method.

Systematic experiments are conducted to simulate the task of drug–target interactions prediction cross four datasets. Compared with state-of-the-art approaches, our proposed approach achieves consistent improvement in terms of area under ROC (AUC) curve and area under precision versus recall (AUPR) curve. As these four datasets contain different portions of new drug candidates and target candidates in the simulation, the improvements of BLM–NII compared with BLM are also different for the four datasets. The most significant improvement is achieved on the nuclear receptor dataset, which contains the largest portion of new candidates. This shows that the NII strategy, i.e. to infer label information or training data from neighbors when there is no training data readily available from the query compound/protein itself, is feasible and effective for dealing with the *new candidate* problem of the original BLM.

2 METHODS

2.1 Problem formalization

Assume that the bipartite interaction network $\mathbf{N}_1$ illustrated in Figure 1 involves m_d drugs/compounds and m_t targets, which are referred to as existing drug candidates and target candidates, respectively. We use matrix $\mathbf{A}$ to represent this network, i.e. $a_{ij} \in \mathbf{A} = 1$ if the i-th compound d_i is known to interact with the j-th target t_j. All other entries of $\mathbf{A}$ are 0. The problem under consideration is how to make use of the known interactions together with the compound similarities and protein similarities to predict new interactions between n_d drug-candidate compounds and n_t target-candidate proteins, where $n_d > m_d$ and $n_t > m_t$. This means there are $\overline{m}_d = n_d - m_d$ new drug candidates and $\overline{m}_t = n_t - m_t$ new target candidates, which have no interactions currently known. The whole network involving n_d compounds and n_t proteins can be represented as

$$\mathrm{N}_{n_d \times n_t} = \left[\begin{array}{c|c} (\mathrm{N}_1)_{m_d \times m_t} & (\mathrm{N}_2)_{m_d \times \overline{m}_t} \\ \hline (\mathrm{N}_3)_{\overline{m}_d \times m_d} & (\mathrm{N}_4)_{\overline{m}_d \times \overline{m}_t} \end{array}\right] = \left[\begin{array}{c|c} \mathbf{A} & 0 \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right], \quad (1)$$

where known interactions correspond to non-zero entries of $\mathbf{A}$. Now, we want to predict possible interactions in $\mathbf{N}_1$ between existing drug candidates and target candidates, as well as in other three subnetworks $\mathbf{N}_2$, $\mathbf{N}_3$ and $\mathbf{N}_4$, where the interactions at least involve one type of new

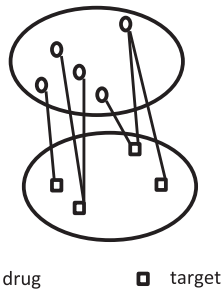

Fig. 1. Bipartite interaction network: a network consists of two types of nodes, where edges only connect different types of nodes. The drug–target interaction network is a bipartite network, where drug and target are two types of nodes and the interactions between them are the edges

candidates, i.e. the target candidate is new, the drug candidate is new or both are new.

2.2 Bipartite local model

To predict p_{ij}, the probability that a drug d_i and a target t_j interact, the basic BLM proposed by Bleakley and Yamanishi (2009) is described as follows. A local model for d_i denoted as $\mathrm{Mod_d}(i)$ is first learned based on its interaction profile $\mathbf{a}'_i$ and the similarities between targets $\mathbf{S}^{\mathrm{t}}$, i.e.

$$\mathrm{Mod_d}(i) = \mathrm{train}(\mathbf{S}^{\mathrm{t}}, \mathbf{a}'_i). \tag{2}$$

Here, train represents the learning process of a certain classifier, e.g. support vector machine or (Kernel) regularized least squares (RLS), the similarity matrix $\mathbf{S}^{\mathrm{t}}$ is used as the observed data of target candidates, and the interaction profile $\mathbf{a}'_i$, i.e. the i-th row vector of $\mathbf{A}$, serves as label information to label each target candidate whether interacting with this drug. Once the model $\mathrm{Mod_d}(i)$ is learned, it is used to predict p^d_{ij}, the probability of interaction between d_i and the query target candidate t_j:

$$p^{\mathrm{d}}_{ij} = \mathrm{test}(\mathrm{Mod_d}(i), \mathbf{s}^{\mathrm{t}}_j), \tag{3}$$

where $\mathbf{s}^t_j$ is the j-th column of $\mathbf{S}^{\mathrm{t}}$ recording the similarities between t_j and other targets. The similar model learning and prediction process are performed independently from the query-target side to get p^t_{ij}, i.e.

$$\mathrm{Mod_t}(j) = \mathrm{train}(\mathbf{S}^{\mathrm{d}}, \mathbf{a}_j), \tag{4}$$

$$p^t_{ij} = \mathrm{test}(\mathrm{Mod_t}(j), \mathbf{s}^d_i), \tag{5}$$

where $\mathbf{a}_j$ is the j-th column vector of $\mathbf{A}$ or the interaction profile of target t_j. Once both p^{d}_{ij} and p^{t}_{ij} have been calculated, they are combined to get probability p_{ij}:

$$p_{ij} = g(p^{\mathrm{d}}_{ij}, p^{\mathrm{t}}_{ij}), \tag{6}$$

where g is a function that combines or integrates p^{d}_{ij} and p^{t}_{ij}. Examples include $p_{ij} = \max\{p^{\mathrm{d}}_{ij}, p^{\mathrm{t}}_{ij}\}$ and $p_{ij} = 0.5(p^{\mathrm{d}}_{ij} + p^{\mathrm{t}}_{ij})$, where g is the *max* or *average* function.

After p_{ij} is calculated for each pair of compound i and protein j, the output network of BLM may be represented as

$$\mathbf{N}_{\mathrm{BLM}} = \left[\begin{array}{c|c} \mathbf{N}^{\mathrm{BLM}}_1 & \mathbf{N}^{\mathrm{BLM}}_2 \\ \hline \mathbf{N}^{\mathrm{BLM}}_3 & 0 \end{array}\right], \tag{7}$$

with

$$\mathbf{N}^{\mathrm{BLM}}_1 = \mathbf{N}_1 + \mathbf{P}_1(\mathrm{Mod_d}, \mathrm{Mod_t}), \tag{8}$$

$$\mathbf{N}^{\mathrm{BLM}}_2 = \mathbf{P}_2(\mathrm{Mod_d}), \tag{9}$$

$$\mathbf{N}^{\mathrm{BLM}}_3 = \mathbf{P}_3(\mathrm{Mod_t}). \tag{10}$$

where $\mathbf{P}_1$ gives the predicted interactions between existing drug candidates and existing target candidates, $\mathbf{P}_2$ are predicted interactions between existing drug candidates and new target candidates and $\mathbf{P}_3$ gives predicted interactions between new drug candidates and existing target candidates.

For any classifier that is used, the known targets of d_i corresponding to non-zero elements of $\mathbf{a}'_i$ and the pairwise target similarity $\mathbf{S}^{\mathrm{t}}$ are critical to the final prediction of p^{d}_{ij}. The model learned for d_i describes how this drug selects targets. Once the model is learned, the similarities between the query target and those known targets of d_i largely decide p^{d}_{ij}. Similarly, known targeting drugs of t_j or non-zero elements of t_j's interaction profile $\mathbf{a}_j$ and the pairwise drug similarity $\mathbf{S}^{\mathrm{d}}$ are critical to the final prediction of p^{t}_{ij}. Under the same BLM framework, different results are produced due to the differences in $\mathbf{S}^{\mathrm{d}}$, $\mathbf{S}^{\mathrm{t}}$, the classifier and the combination function g. According to the study of Laarhoven *et al.* (2011), network-based similarity which encodes the topology information of the interaction network has been shown to provide good results. With the Gaussian kernel, the network-based drug similarity $\mathbf{S}^{\mathrm{d}}_n$ and network-based target similarity $\mathbf{S}^{\mathrm{t}}_n$ are calculated as:

$$\mathbf{S}^{\mathrm{d}}_n(i,j) = \exp\Big(-\frac{\|\mathbf{a}'_i - \mathbf{a}'_j\|^2}{\gamma}\Big), \tag{11}$$

$$\mathbf{S}^{\mathrm{t}}_n(i,j) = \exp\Big(-\frac{\|\mathbf{a}_i - \mathbf{a}_j\|^2}{\gamma}\Big), \tag{12}$$

where the bandwidth $\gamma = \gamma_0 * \frac{1}{n}\sum_{i=1}^{n} a^2_{ij}$, and different bandwidths may be used for drug and target, respectively. However, the result with network-based similarity may not remain good when the information contained in the interaction network is not sufficient enough. Rather than considering one type of similarity, a more general way is to combine several types of similarities. Here, we use both the network-based similarity and chemical similarity for drug similarity $\mathbf{S}^{\mathrm{d}}$, and the network-based similarity and sequence similarity for target similarity $\mathbf{S}^{\mathrm{t}}$ through linear combination:

$$\mathbf{S}^{\mathrm{d}} = \alpha\mathbf{S}^{\mathrm{d}}_c + (1-\alpha)\mathbf{S}^{\mathrm{d}}_n, \tag{13}$$

$$\mathbf{S}^{\mathrm{t}} = \alpha\mathbf{S}^{\mathrm{t}}_s + (1-\alpha)\mathbf{S}^{\mathrm{t}}_n, \tag{14}$$

where $\mathbf{S}^d_c$ is the chemical structure similarity for drug, $\mathbf{S}^t_s$ is the amino acid sequence similarity for protein and α is the combination weight set by user. Although more sophisticated ways such as Kronecker product may be used to combine two types of similarity matrices or kernel matrices, experimental results in (Laarhoven *et al.* 2011) show that the linear combination gives comparable performance with a much lower computational complexity.

2.3 Neighbor-based interaction-profile inferring

Good performance of supervised learning is largely dependent on the amount and quality of labeled training data. When a drug/target candidate is new, it has no existing interactions that can be used as label information and the model for this candidate thus can not be learned. As shown in (7), interactions between new drug candidates and new target candidates remain unpredicted in BLM. To extend the application domain of BLM to new drug/target candidates, we propose to derive training data from their neighbors. Based on the assumption that drugs/compounds which are similar to each other interact with the same targets, interaction profile for new drug-candidate compounds could be possibly inferred from their neighbors' interactions. Compounds with large similarities to the new drug-candidate compound are said to be its neighbors. Since new drug-candidate compounds have no interactions, or all the elements of its current interaction profile vector are 0, it is not suitable to consider network-based similarity here, so only chemical structure similarity is used to define the neighbors of a drug-candidate compound. Formally, for a compound d_i which is a new drug-candidate, we infer the j-th dimension of its interaction profile $\mathbf{l}^{\mathrm{d}}(i)$ with

$$l^{\mathrm{d}}_j(i) = \sum_{h=1}^{m_{\mathrm{d}}} s_{ih}a_{hj}, \tag{15}$$

where s_{ih} is the chemical similarity between two compounds d_i and d_h. The above formula shows that the interaction weight of this drug with respect to the j-th target is the collection of its neighbors' interactions to this target. For a given new drug-candidate compound, the simple formula given in Equation (15) defines that the inferred weight of interaction between this compound and a target is high if many of its neighbors interact with this target, and also it is decided more by neighbors with large similarities than those with small similarities. Since new target-candidate proteins have no interactions with any compound, the

inferred interactions for d_i are only with existing target candidates. To be more specific, $I_j^{\mathrm{d}}(i) > 0$ if the j-th target candidate is an existing one, i.e. $a_{hj} > 0$ for at least one h, and $I_j^{\mathrm{d}}(i) = 0$ if the j-th target candidate is new, i.e. $a_{hj} = 0$ for all h. To ensure the value of each $I_j^{\mathrm{d}}(i)$ is in the range of [0, 1], linear scale is performed subsequently, i.e. $I_j^{\mathrm{d}}(i) = (I_j^{\mathrm{d}}(i) - \min_h I_h^{\mathrm{d}}(i))/(\max_h I_h^{\mathrm{d}}(i) - \min_h I_h^{\mathrm{d}}(i))$. After we obtained the inferred interaction profile, we can use it as label information to learn the model of d_i:

$$\mathrm{Mod}'_{\mathrm{d}}(\mathrm{i}) = \mathrm{train}(\mathbf{S}^{\mathrm{t}}, \mathbf{I}^{\mathrm{d}}(i)). \tag{16}$$

In the same way, this procedure is applied to a new target-candidate protein t_j to obtain its inferred interaction profile $\mathbf{I}^{\mathrm{t}}(j)$, where its neighbors are defined based on sequence similarity. The model of t_j can then be learned with $\mathbf{I}^{\mathrm{t}}(j)$:

$$\mathrm{Mod}'_{\mathrm{t}}(\mathrm{j}) = \mathrm{train}(\mathbf{S}^{\mathrm{d}}, \mathbf{I}^{\mathrm{t}}(j)). \tag{17}$$

This interaction profile inferring technique is particularly useful for those new drug/target candidates, for which existing supervised methods (e.g. BLM) fail to produce reasonable predictions. It can also be useful to enhance the classification models for any compounds/proteins without enough training data or label information.

2.4 BLM with NII

By integrating the above presented NII strategy into the BLM framework, we have the BLM with NII (BLM–NII). The detailed steps of BLM–NII to predict the probability p_{ij} between any compound i and any protein j is described in Algorithms 1 and 2.

Algorithm 1: BLM–NII

input : $\mathbf{A}$, $\mathbf{S}_c^d$, $\mathbf{S}_s^t$
output: p_{ij}
get p_{ij}^d = NII-integrated Learning and Prediction ($\mathbf{A}$, $\mathbf{S}_c^d$, $\mathbf{S}_s^t$) from d_i;
get p_{ij}^t = NII-integrated Learning and Prediction ($\mathbf{A}$, $\mathbf{S}_s^t$, $\mathbf{S}_c^d$) from t_j;
Combine p_{ij}^d and p_{ij}^t to get the final result $p_{ij} = g(p_{ij}^d, p_{ij}^t)$

Algorithm 2: NII-integrated learning and prediction

input : $\mathbf{A}$, $\mathbf{S}_c^d$, $\mathbf{S}_s^t$
output: p_{ij}^d
if d_i *is new* **then**
  obtain $\mathbf{I}^d(i)$ with Eq. (15) with $\mathbf{S}_c^d$
else
  $\mathbf{I}^d(i)$ is the i-th row of $\mathbf{A}$
end
Compute $\mathbf{S}_n^t$ with Eq. (12) and $\mathbf{S}^t$ with Eq. (14);
Learn a local model for d_i, i.e., $Mod_d(i) = train(\mathbf{S}^t, \mathbf{I}^d(i))$;
if t_j *is new* **then**
  predict p_{ij}^d with $Mod_d(i)$ and $\mathbf{S}_s^t$
else
  predict p_{ij}^d with $Mod_d(i)$ and $\mathbf{S}^t$
end

The output network of BLM–NII is expressed as

$$\mathbf{N}_{\mathrm{BLM-NII}} = \left[\begin{array}{c|c} \mathbf{N}_1^{\mathrm{BLM-NII}} & \mathbf{N}_2^{\mathrm{BLM-NII}} \\ \hline \mathbf{N}_3^{\mathrm{BLM-NII}} & \mathbf{N}_4^{\mathrm{BLM-NII}} \end{array}\right], \tag{18}$$

with

$$\mathbf{N}_1^{\mathrm{BLM-NII}} = \mathbf{N}_1^{\mathrm{BLM}}, \tag{19}$$

$$\mathbf{N}_2^{\mathrm{BLM-NII}} = \mathbf{P}_2(\mathrm{Mod}_{\mathrm{d}}, \mathrm{Mod}'_{\mathrm{t}}), \tag{20}$$

$$\mathbf{N}_3^{\mathrm{BLM-NII}} = \mathbf{P}_3(\mathrm{Mod}'_{\mathrm{d}}, \mathrm{Mod}_{\mathrm{t}}), \tag{21}$$

$$\mathbf{N}_4^{\mathrm{BLM-NII}} = \mathbf{P}_4(\mathrm{Mod}'_{\mathrm{d}}, \mathrm{Mod}'_{\mathrm{t}}). \tag{22}$$

Comparing $\mathbf{N}_{\mathrm{BLM-NII}}$ and $\mathbf{N}_{\mathrm{BLM}}$, it is observed that the interactions between existing drug candiates and target candiates are the same for the two approaches, while the interactions in the other three cases in BLM–NII are different from those in BLM. First, BLM–NII is able to predict $\mathbf{P}_4$, the interactions between drug candidates and target candidates that are both new. Second, $\mathbf{P}_2$ and $\mathbf{P}_3$ in BLM–NII are predicted from both the drug side and the target side, while in BLM are predicted only from one side.

Learning from neighbors allows drug/target candidates to obtain labeled data when themselves do not have or have insufficient labeled data for training. This procedure actually introduces some degree of globalization into the original local model to provide more chances of learning from known knowledge. However, too much globalization is not desired as it could eliminate the local characteristics and make the models of individual candidates less discriminative. Moreover, the low quality of neighbors due to imprecise similarity measure may cause negative impact when the learning process replies on too much neighbors' information. In other words, the inferred interaction profile, although is helpful, may introduce a certain amount of noise. Therefore, in this study, we only activate the neighbor-based learning for totally new candidates. For other cases, we still train the model locally with its own known interactions.

3 MATERIALS

To facilitate comparison with published approaches, we used the same groups of four datasets which are first analyzed by Yamanishi *et al.* (2008) and then later by Bleakley and Yamanishi (2009), Xia *et al.* (2010), Laarhoven *et al.* (2011) and Cheng *et al.* (2012). These four datasets correspond to drug–target interactions of four important categories of protein targets, namely enzyme, ion channel, G-protein-coupled receptor (GPCR) and nuclear receptor, respectively. The datasets were downloaded from http://web.kuicr.kyoto-u.ac.jp/supp/yoshi/drugtarget/.

Table 1 gives some statistics of each dataset including the total number of drugs (n_{d}), the total number of targets (n_{t}), the total number of interactions (E), the average number of targets for each drug ($\bar{D}_{\mathrm{d}}$), the average number of targeting drugs for each target ($\bar{D}_{\mathrm{t}}$), the percentage of drugs that have only one target ($D_{\mathrm{d}} = 1$) and the percentage of targets that have one targeting drug ($D_{\mathrm{t}} = 1$). It is shown from this table that among the four drug–target interaction networks, on average, each drug and target in ion channel and enzyme have more interactions than those in GPCR and nuclear receptor. It is also worthy noting that in the leave-one-out cross–validation (LOOCV), drugs and targets with one interaction are 'new candidates' as the only one interaction is covered over to leave no recorded interaction, e.g. 72% drugs in the nuclear receptor are 'new candidates' in the simulation.

Each dataset is described by three types of information in the form of three matrices: (i) the drug–target interaction matrix; (ii) the drug–drug similarity matrix and (iii) the target–target similarity matrix. The interaction networks were retrieved from the KEGG BRITE (Kanehisa *et al.*, 2006), BRENDA (Schomburg

Table 1. Some statistics of the four datasets

Dataset	Enzyme	Ion channel	GPCR	Nuclear receptor
n_d	445	210	223	54
n_t	664	204	95	26
E	2926	1476	635	90
$\bar{D_d}$	6.58	7.03	2.85	1.67
$\bar{D_t}$	4.41	7.24	6.68	3.46
$D_d = 1(\%)$	39.78	38.57	47.53	72.22
$D_t = 1(\%)$	43.37	11.27	35.79	30.77

Table 2. Comparison with existing approaches for the four datasets

Dataset	Method	AUC	AUPR
Enzyme	Weighted profile	86.4	6.30
	BY(2009)	97.6	83.3
	Laarhoven *et al*. (2011)	97.8	91.5
	BLM–NII	98.8	92.9
Ion channel	Weighted profile	81.9	17.2
	BY(2009)	97.3	78.1
	Laarhoven *et al*. (2011)	98.4	94.3
	BLM–NII	99.0	95.0
GPCR	Weighted profile	76.5	10.9
	BY(2009)	95.5	66.7
	Laarhoven *et al*. (2011)	95.4	79.0
	BLM–NII	98.4	86.5
Nuclear receptor	Weighted profile	74.9	17.1
	BY(2009)	88.1	61.2
	Laarhoven *et al*. (2011)	92.2	68.4
	BLM–NII	98.1	86.6

et al., 2004), SuperTarget (Gnther *et al*., 2008) and DrugBank (Wishart *et al*., 2008). The drug–drug similarity is measured based on chemical structures from the DRUG and COMPOUND sections in the KEGG LIGAND database (Kanehisa *et al*., 2006). The chemical structure similarities between drugs are computed with SIMCOMP (Hattori *et al*., 2003), which uses a graph alignment algorithm to get a global similarity score based on the size of the common substructures between two compounds. The target–target similarity is measured based on the amino acid sequences retrieved from the KEGG GENES database (Kanehisa *et al*., 2006). The sequence similarities between proteins are computed with a normalized version of Smith–Waterman score. More details on how the data have been collected and calculated are given in Yamanishi *et al*. (2008).

4 EVALUATION

Systematic experiments are performed to evaluate the performance of the presented approach with datasets summarized in Table 1. As in Laarhoven *et al*. (2011), LOOCV is performed. Since the real interaction to be predicted is left out, compounds and proteins with one interaction (i.e. $D_d = 1$ or $D_t = 1$) turn out to have no training data and thus they are treated as 'new candidates' in the cross-validation. To test the robustness of the presented approach, we also performed 10-fold cross-validation. The results of 10 trials 10-fold cross-validation can be found in Tables S5–S8 of the Supplementary Material.

4.1 Compare with state-of-the-art approaches

First, we compare the performance of BLM–NII ($g = \max$, $\alpha = 0.5$) with the weighted profile method (Yamanishi *et al*., 2008) and two other state-of-the-art approaches (Bleakley and Yamanishi, 2009) and (Laarhoven *et al*., 2011) denoted as BY (2009) and Laarhoven *et al*. (2011), respectively. The same RLS classifier is used for BLM–NII as Laarhoven *et al*. (2011). We measure the quality of the predicted interactions in terms of AUC curve (or true-positive rate versus false-positive rate curve) and AUPR curve.

Table 2 gives the AUC and AUPR scores of the four approaches for the four datasets. The results of BY (2009) and Laarhoven *et al*. (2011) are the best ones reported in Bleakley and Yamanishi (2009) and Laarhoven *et al*. (2011), respectively. From this table, it is clear that BLM–NII outperforms the other three for all the datasets. Since the results of weighted profile are much worse than those of the three BLM-based methods namely BY (2009), Laarhoven *et al*. (2011) and BLM–NII, we now focus on the comparison of these three approaches. As been discussed in Laarhoven *et al*. (2011), by incorporating the network-based similarity, the performance of BLM can be improved, i.e. the results of Laarhoven *et al*. (2011) in terms of AUPR are much better than those of BY (2009). It is also shown that the performance of BLM can further be improved by integrating the NII procedure, i.e. the results of BLM–NII is consistently better than those of Laarhoven *et al*. (2011).

It is interesting to observe that different levels of improvements have been achieved for different datasets. Comparing Laarhoven *et al*. (2011) and BY (2009), the improvement is the most significant on ion channel and the least significant on nuclear receptor. Differently, comparing BLM–NII and Laarhoven *et al*. (2011), the improvement is the largest for nuclear receptor and the least for ion channel. Such kind of differences are expected due to the differences in the structure of the datasets. From Table 1, it is shown that among the four datasets, the average numbers of interactions of each drug and target are the largest for ion channel and the smallest for nuclear receptor. This means that the interaction network of ion channel contains more information than nuclear receptor and thus the network-based similarity of ion channel is more robust and informative than that of nuclear receptor. Therefore, incorporating the network-based similarity results in larger improvement for ion channel. Since drugs or targets with one interaction are 'new candidates' in the simulation, it is also shown from Table 1 that the nuclear receptor contains the largest portion of 'new candidates' while the ion channel contains the least. Thus, by applying the NII procedure, BLM–NII has more chances to improve the results for nuclear receptor than for Ion Channel.

4.2 Comparison between BLM and BLM–NII

To directly show the improvements attributed to the NII strategy, we now compare BLM–NII and BLM, i.e. the results of

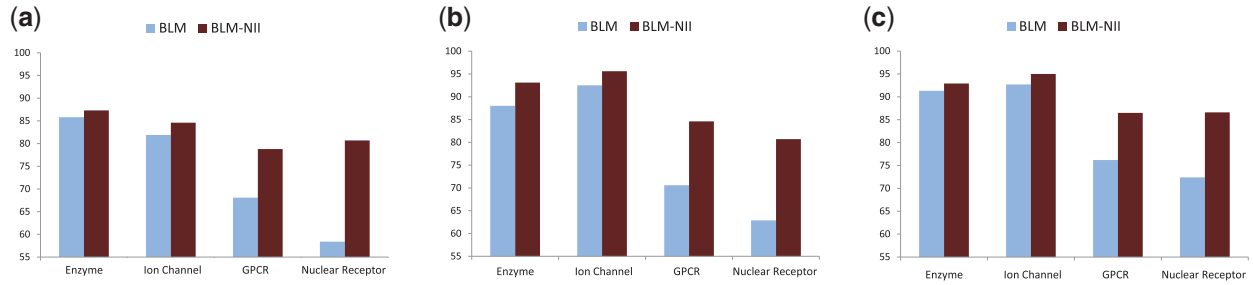

Fig. 2. AUPR of BLM and BLM–NII for nuclear receptor with different types of similarities: (a) $\alpha = 1$, (b) $\alpha = 0$ and (c) $\alpha = 0.5$

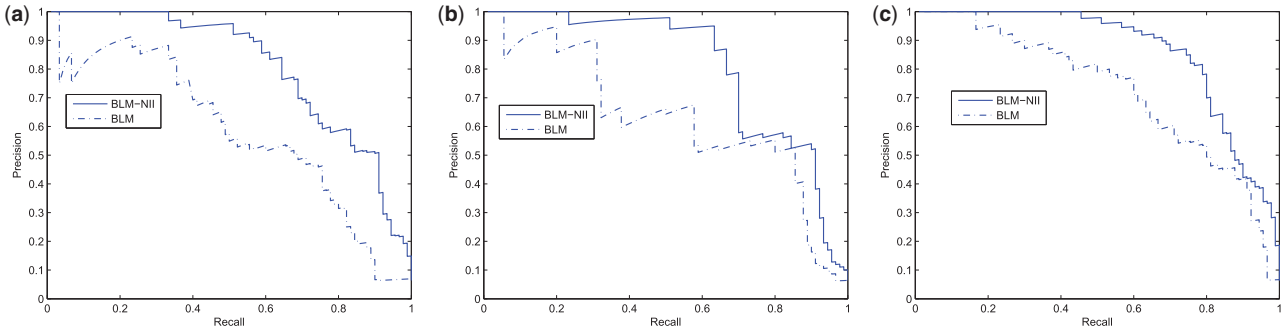

Fig. 3. Precision–recall curve of BLM and BLM–NII for nuclear receptor with different similarities: (a) $\alpha = 1$, (b) $\alpha = 0$ and (c) $\alpha = 0.5$

Table 3. Compare 1% and 3% top ranked pairs of BLM and BLM–NII for nuclear receptor

		Top 1%			Top 3%		
α	Method	Sensitivity	PPV	MCC	Sensitivity	PPV	MCC
1	BLM	13.3	85.7	32.5	35.6	76.2	50.0
	BLM–NII	15.6	100.0	38.3	44.4	95.2	63.7
0	BLM	16.7	93.8	38.3	32.2	67.4	44.3
	BLM–NII	18.9	100.0	42.3	45.6	97.6	65.4
0.5	BLM	15.6	100.0	38.3	40.0	85.7	56.9
	BLM–NII	15.6	100.0	38.3	45.6	97.6	65.4

BLM–NII where new candidates are treated as existing ones. We applied both BLM and BLM–NII with three different groups of inputs by setting α in Equations (13) and (14) to 1, 0 and 0.5.

We obtained the AUC and AUPR scores of both methods with $g = \max$. The results of both with $g =$ average or mean have also been produced, which can be found in Tables S1–S4 of the Supplementary Material. Since the same conclusion can be drawn with respect to either of the two metrics, we put the AUC scores in the Supplementary Material and plot the AUPR scores of BLM and BLM–NII for the four datasets with three different types of similarities in Figure 2. It is shown that for any type of similarities, BLM–NII performs better than BLM for all the datasets. Again, the improvements made by BLM–NII are more significant for nuclear receptor and GPCR than for the other two datasets.

Now using nuclear receptor, we make further comparison of the performance between BLM and BLM–NII. Figure 3 plots the precision–recall curve of BLM and BLM–NII. Table 3 shows the sensitivity (or recall), PPV (positive predictive value or precision) and MCC (Matthews correlation coefficient). The two groups of results in Table 3 are calculated by considering the 1% and 3% pairs with the highest p_{ij} values as positive, respectively. It is clearly shown from these results that with NII being integrated, the performance of BLM has been improved.

4.3 Detailed analysis of the effectiveness of NII

To take a close look at the difference in the results attributed to the NII strategy, we now compare those top ranked interactions of the nuclear receptor dataset produced by BLM–NII and BLM. Since this dataset has 90 known interactions, we inspect the 90 interactions with the highest probabilities predicted by each algorithm.

As summarized in Table 4 (More detailed results are in Table S11 of the Supplementary Material), among the top 90 predicted interactions, BLM only correctly detected 58 known interactions while BLM–NII detected 71, and 57 known interactions are ranked within 90 by both. Although one interaction detected by BLM is missed by BLM–NII, this one ranks 104 in BLM–NII, which indicates that this pair is still recognized to be interacting with a highly possibility by BLM–NII. Nevertheless, 14 interactions detected by BLM–NII are missed by BLM. The average rank of these 14 interactions produced by BLM is 388 as some of them ranks very low.

Among these 14 drug–target pairs, three pairs namely D00163 (Chenodeoxycholic acid) – hsa9971 (nuclear receptor subfamily 1, group H, member 4), D00506 (Phenobarbital) – hsa9970 (nuclear receptor subfamily 1, group I, member 3) and D05341 [Palmitic acid (NF)] – hsa3174 (hepatocyte nuclear factor 4, gamma), which are assigned extremely low ranks by BLM are successfully detected by BLM–NII as shown in Figure 4. After checking, we find that the query drug D00163 of the first pair only has one target which happens to be the query target hsa9971, and the query target is known to be only interacting with the query drug. The other two pairs have the same situation as this pair. As we left out the true interaction in our simulation, the testing for these three pairs becomes to predict interaction between new drug-candidate compound and new target-candidate protein. Since training data are absent for both the query drug and query target, BLM fails to detect interactions for those three pairs. Although difficulty is presented for such kind of cases, BLM–NII successfully detected these three pairs to be interacting. This shows the effectiveness of NII for prediction of interaction involving new candidates.

Table 4. Performance of BLM and BLM–NII on nuclear receptor

Total known interactions:	90
Interactions detected by BLM	58
Interactions detected by BLM–NII	71
Interactions detected by both BLM and BLM-–NII	57

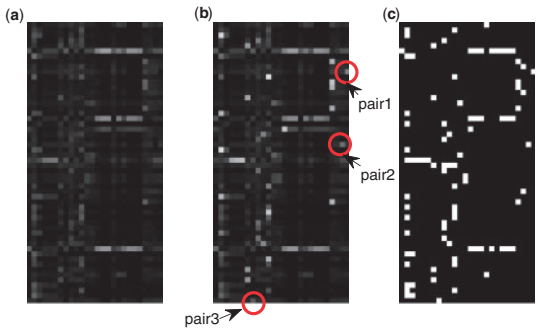

Fig. 4. Drug–target interaction matrix of nuclear receptor. (**a**) Predicted by BLM, (**b**) predicted by BLM–NII, (**c**) real interaction matrix. Each entry of the interaction matrix is plotted as a pixel. The brightness of a pixel represents the interaction possibility of the corresponding pair, i.e. the brighter the more possible that the pair interacts. Three pairs are circled in (b). These three pairs which consist of drug candidate and protein candidate that are both 'new' in Loo validation are detected by BLM–NII

Now using D00163 and hsa9971 as an example, we give intermediate results to illustrate how NII helps detect the interactions between new drug-candidate compounds and new target-candidate proteins. Figure 5 shows the local model learned for D00163 with the help of inferred training data. Specifically, Figure 5a shows the inferred interaction profile of D00163, i.e. the weighted interactions between D00163 and 25 non-query targets calculated with Equation (15). It shows that the associations between D00163 and several targets such as hsa2099 are large. This is because many of D00163's neighbors or similar drugs, such as D00066, interact with this target as seen from Figure 5b. Using this inferred interaction profile as label information, Figure 5c shows the learned local model of D00163, or the weight of each of the targets learned with the classifier with respect to D00163. With this learned model, BLM–NII successfully detected the interaction between D00163 and hsa9971 based on the similarities between the query target hsa9971 and other targets especially those with large weights in the model of D00163. In the same manner, the local model of the query target which is a 'new' candidate can be learned with NII. This example illustrates the feasibility and effectiveness of the presented approach to infer training data or label information from the interaction profiles of neighbors.

5 CONCLUSION AND DISCUSSION

We proposed an intuitive solution to the *new candidate* problem of BLM by integrating a NII procedure, i.e. infer training data

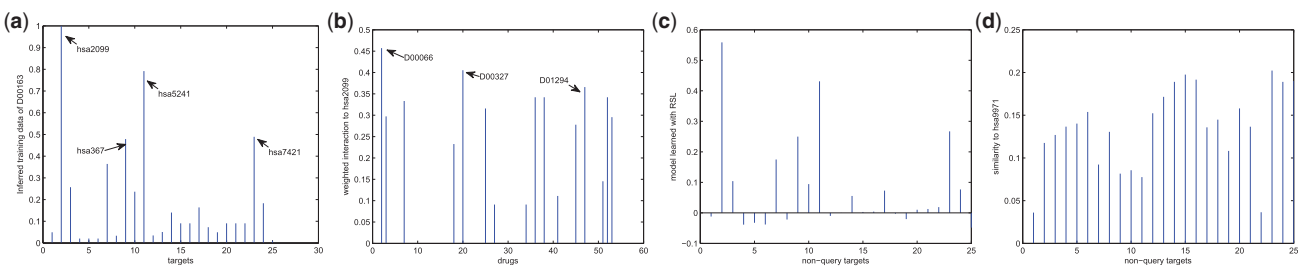

Fig. 5. Local model learning for D00163 with BLM–NII. (a) inferred interaction profile $\mathbf{I}^d$ of D00163, (b) weighted interaction of D00163's neighbors to hsa2099 calculated with $s(D00163, i) \times a(i, hsa2099)$ for each drug i, (c) learned model of D00163 by the RLS classifier, (d) similarities between hsa9971 and other proteins, i.e. $\mathbf{s}^t_{has9971}$

from neighbors' interaction profiles. Through systematic experiments with benchmark datasets, we demonstrated the effectiveness of BLM–NII for predicting interactions between new drug-candidate compounds and new target-candidate proteins.

In the presented approach, we allow all the neighbors to participate in training data inferring. To allow only neighbors with large similarities to contribute, a threshold may be used to reduce the impact of those non-important neighbors to 0. Alternately, a Gaussian function may be introduced to gradually decrease the influence of neighbors based on their distances to the new drug/target candidate in query.

In the current work, we only apply the NII procedure for those completely new candidates that have no existing training data at all, and we find that the results are already good enough to show the usefulness of NII. Since it is quite common that drugs only activate or inhibit a small number of targets and targets are only activated or inhibited by very limited drugs, the NII procedure may be applied to drugs and targets which do not have sufficient training data. We expect that more accurate prediction models may be build by using neighbors' information to enhance the limited training examples. However, too much emphasis on neighbors tends to eliminate the local characteristics of each drug and target and could cause deterioration in the prediction performance. Nevertheless, it would be an interesting future work to explore the balance between local information and global information in model learning.

Funding: This research was supported by Singapore MOE AcRF (MOE2008-T2-1-074) and Startup (M4080108.020) from Nanyang Technological University, Singapore.

Conflict of Interest: none declared.

REFERENCES

Bleakley,K. and Yamanishi,Y. (2009) Supervised prediction of drug–target interactions using bipartite local models. *Bioinformatics*, **25**, 2397–2403.

Campillos,M. *et al.* (2008) Drug target identification using side-effect similarity. *Science*, **321**, 263–266.

Chen,X. *et al.* (2012) Drug–target interaction prediction by random walk on the heterogeneous network. *Mol. BioSyst*, **8**, 1970–1978.

Cheng,F. *et al.* (2012) Prediction of drug–target interactions and drug repositioning via network-based inference. *PLoS Comput. Biol.*, **8**, e1002503.

Gunther,S. *et al.* (2008) Supertarget and matador: resources for exploring drug–target relationships. *Nucleic Acids Res.*, **36**, D919–D922.

Hattori,M. *et al.* (2003) Development of a chemical structure comparison method for integrated analysis of chemical and genomic information in the metabolic pathways. *J. Am. Chem Soc.*, **125**, 11853–11865.

Haupt,V.J. and Schroeder,M. (2011) Old friends in new guise: repositioning of known drugs with structural bioinformatics. *Breif. Bioinform*, **12**, 312–326.

He,Z. *et al.* (2010) Predicting drug–target interaction networks based on functional groups and biological features. *PLoS One*, **5**, e9603.

Kanehisa,M. *et al.* (2006) From genomics to chemical genomics: new developments in kegg. *Nucleic Acids Res.*, **34**, D354–D357.

Keiser,M.J. *et al.* (2009) Predicting new molecular targets for known drugs. *Nature*, **462**, 175–181.

Laarhoven,T.V. *et al.* (2011) Gaussian interaction profile kernels for predicting drug–target interaction. *Bioinformatics*, **27**, 3036–3043.

Laggner,C. *et al.* (2012) Chemical informatics and target identification in a zebrafish phenotypic screen. *Nat. Chem. Biol.*, **8**, 144–146.

Martin,Y.C. *et al.* (2002) Do structurally similar molecules have similar biological activity? *J. Med. Chem.*, **45**, 4350–4358.

Perlman,L. *et al.* (2011) Combining durg and gene similarity measures for drug–target elucidation. *J. Comput. Biol.*, **18**, 133–145.

Schomburg,I. *et al.* (2004) Brenda, the enzyme database: updates and major new developments. *Nucleic Acids Res.*, **32**, D431–D433.

Wishart,D.S. *et al.* (2008) Drugbank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res.*, **36**, D901–D906.

Xia,Z. *et al.* (2010) Semi-supervised drug–protein interaction prediction from heterogeneous biological spaces. *BMC Syst. Biol.*, **4** (**Suppl. 2**), S6.

Yamanishi,Y. *et al.* (2008) Prediction of drug–target interaction networks from the integration of chemical and genomic spaces. *Bioinformatics*, **24**, i232–i240.